情報処理学会論文誌 Vol.63 No.2 615–623 (Feb. 2022)

人の移動データに基づく地域のクラスタリングと入込観光客数予測への応用

落合 桂一[1,a)] 寺田 雅之[1]

受付日 2021年4月6日, 採録日 2021年11月2日

概要：人々の社会生活において移動は基本的な行動であり，人々の移動は経済活動や交通，公衆衛生など様々な分野と関わりが深い．人々の移動データと応用先の分野のデータを組み合わせて解析することで，経済活動の予測や交通の最適化など各分野での分析精度の向上や，より詳細な分析が期待される．そこで本研究では，人の移動データに基づいて市町村をクラスタリングすることの有用性を実データを使って評価する．まず，約 11 万ユーザの半年間の移動データから市町村をクラスタリングする．次に，クラスタリングの有用性を，(1) 市町村別観光消費額の年間推移の類似性，(2) 入込観光客数の予測問題という 2 つの評価実験により検証する．予測問題では，行政が決めた地域区分と比較して予測誤差を削減できることを確認した．

キーワード：ネットワーク分析，コミュニティ検出，教師あり機械学習

Location Clustering based on Human Mobility and Its Application for Prediction of Sightseeing Visitors

Keiichi Ochiai[1,a)] Masayuki Terada[1]

Received: April 6, 2021, Accepted: November 2, 2021

Abstract: In the social life of the people, the movement is a fundamental action, and the movement of the people is deeply related to various fields such as economic activity, traffic, and public health. By combining the movement data of the people and the data of the application field such as the economic activity and optimization of the traffic, it is expected that the analysis accuracy can be improved and more detailed analysis can be conducted. In this study, we evaluate the usefulness of clustering cities based on human movement data using real data. At first, cities are clustered from mobility data of about 110,000 users for half a year. Next, the usefulness of clustering is verified by two evaluation experiments: (1) similarity of annual transition of sightseeing consumption by municipalities, and (2) prediction problem of the number of entering tourists. In the prediction problem, it was confirmed that the prediction error could be reduced in comparison with the region division decided by the administration.

Keywords: network analysis, community detection, supervised machine learning

1. はじめに

人々の社会生活において移動は基本的な行動であり，人々の移動は経済活動や交通，公衆衛生など様々な分野と関わりが深い．人々の移動データと応用先の分野のデータを組み合わせて解析することで，経済活動の予測や交通の最適化など各分野での分析精度の向上や，より詳細な分析が実施可能になると期待される．たとえば，小売業の販売額を都道府県ごとに予測することを考える．このとき，予測対象の都道府県の過去の販売額推移だけでなく，近隣の都道府県の推移を活用することで予測精度を向上させることが考えられる．具体的には，隣接している都道府県や同一の地域の情報を利用することができる．しかし，隣接している都道府県でも地理的な制約で 2 つの都道府県を結ぶ交通

[1] 株式会社 NTT ドコモ
NTT DOCOMO, INC., Chiyoda-ku, Tokyo 100–6150, Japan
[a)] ochiaike@nttdoomo.com

© 2022 Information Processing Society of Japan

機関が少ないことに起因して人の移動が少ない場合，隣接している都道府県の情報を加味しても予測精度の向上には寄与しないと考えられる．そのため，都道府県の地理的な配置のみではなく，人の移動に基づいて都道府県の結び付きを考慮することが有用であると考えられる．

そこで本研究では，人の移動データに基づいて市町村をクラスタリングし地域を分けることの有用性を実データを使って評価する．本研究では単一または複数の市町村からなる集合を地域と定義する．本研究は，(1) 移動データの収集，(2) 移動データに基づくクラスタリング，(3) 有用性の評価の 3 つのパートから構成される．人の移動データには，Foursquare 社が提供する Swarm [1] というチェックイン共有 SNS でのチェックインデータを利用する．期間は 2018/10/1–2019/3/31 の半年間で，約 11 万ユーザの約 730 万件のチェックインデータを収集した．市町村をノードとし，市町村間の移動があった箇所にエッジを設定することで市町村グラフを構築した．エッジの重みには市町村間を移動したユニークユーザ数を利用する．そして，市町村グラフに対してノードのクラスタリングを行う．グラフのクラスタリング手法の違いによる結果への影響を考察するため，Newman Algorithm [2] および Clauset-Newman-Moore (CNM) Algorithm [3] を利用してクラスタリングを行い結果の比較を行う．最後に，クラスタリングの有用性を観光消費額の推移の類似性，および入込観光客数予測問題で検証する．観光消費額の推移の類似性の評価では，神奈川県の各市町村の観光消費額の推移を，行政が定めた地域区分と移動データに基づく地域区分で比較することで評価する．入込観光客数予測では，神奈川県内の各市町村の入込観光の予測に市町村のクラスタリングの結果を特徴量として利用することで，行政が定めた地域区分との予測精度への影響を比較し評価する．

本研究の貢献は以下のとおりである．

- 大規模な移動データの統計データに基づき地域のクラスタリングを行い，行政により決められた地域区分と移動に基づく地域区分では異なることを明らかにした．
- 移動データに基づくクラスタリングにより，観光消費額の推移がより類似している市町村をクラスタリングできることを定量的に示した．
- 移動データに基づくクラスタリングを地域に関する統計指標の予測に活用することを提案し，神奈川県の入込観光客数の予測問題において有効性を検証した．

本稿の構成は以下のとおりである．次章で人の移動データの分析の既存研究について概観する．次に，3 章で提案手法の詳細を説明し，4 章では実データを用いた評価を行い，既存手法と提案手法の精度について検証する．最後に 5 章で本研究のまとめおよび今後の課題について述べる．

2. 関連研究

人の移動データの分析に関わる研究は (1) 個々人の移動を対象にした研究と，(2) 個人のデータを集約し集団としての移動を分析した研究の 2 つに大きく分けられる．本章ではそれぞれの関連研究についてまとめる．

2.1 個人の移動データの分析

個人の位置情報を解析する研究では，過去の位置情報を元にした将来の目的地予測 [4]，移動手段の推定 [5]，訪問した施設（POI：Point of Interest）の推定 [6] など，様々な研究が行われている．また，過去に訪問した POI を元にユーザの嗜好や場所の特性から POI 推薦を行う研究もさかんに研究されている [7], [8], [9], [10]．これらの研究では，推定モデルの入力が個人の位置情報履歴であり，出力（予測や推定の結果）も個人の位置や移動に関する情報となる点が共通している．

2.2 集団の移動データの分析

本研究とより関連するのは本節で述べる集団の移動に関する分析である．Thiemann ら [11] は，ある場所 i から j への紙幣の移動量が人の移動量と相関するという仮定のもと，wheresgeorge.com というアメリカでの紙幣の場所の追跡ゲームサイトのデータを利用し，人の地域間の移動データを収集した．郵便番号単位の地域分割を行い，各地域をノード，地域間にエッジを設定し，エッジの重みには移動量を利用し人の移動をグラフとして表現した．そして，Modularity に基づくコミュニティ検出手法を適用することで地域のクラスタリングを行った．その結果，行政による地域区分と異なるクラスタが見つかったと報告されている．Rinzivillo ら [12] は，車載 GPS のデータセットを利用して，各位置をノード，移動があった位置どうしにエッジを設定することで移動を表すグラフ構築した．構築したグラフに対して，infomap というネットワークのコミュニティ検出手法を適用することで地域のクラスタリングを行った．移動に基づくクラスタリング結果と行政区分がマッチしていたと報告している．移動データではないが，地域の結び付きを分析した研究に Ratti らの研究がある [13]．Ratti らはイギリス国内における固定電話の発信/着信のデータから，各場所をノード，場所間の結び付きを通話時間としてグラフを構築し，Modularity に基づくコミュニティ検出手法を適用することで地域のクラスタリングを行った．クラスタリングの結果を活用した分析の事例として，イギリスの各地域が独立した場合の影響を分析し，スコットランドが独立した場合の影響が，他の地域が独立した場合と比べて相対的に少ないことを試算した．これらの研究では，移動などの人の生活のデータに基づいて都市をクラスタリングし，行政区分との比較や各地域の国全体への影響を評価

© 2022 Information Processing Society of Japan

している．本研究でも，既存研究と同様に移動に基づくクラスタと行政区分との比較は行うが，日本を対象に移動に基づいて市区町村をクラスタリングした研究は著者らの知る限り行われていない．また，クラスタリング結果を予測問題に応用し評価した研究はない．

3. 移動データに基づく地域のクラスタリング

3.1 データ

人の移動データとして，Foursquare 社が提供するチェックイン共有 SNS である Swarm [1] のチェックインデータを利用する．Swarm の API では自分自身のチェックイン情報は取得可能であるが，プライバシ保護のため他のユーザのチェックイン情報は取得できない．Swarm では，ある場所でチェックインする際，Twitter にチェックイン情報を連携することでツイートとして投稿することができる．本研究では，Twitter API [14] を利用して Swarm のチェックインが Twitter に投稿されたデータを収集した．期間は 2018/10/1–2019/3/31 の半年間で，約 11 万ユーザの約 730 万件のチェックインデータを収集した．

3.2 移動データに基づくグラフ構築

収集したチェックインデータから各ユーザの連続するチェックインをユーザの移動と見なす．実際には Twitter 連携されなかったチェックインや，すべての場所でチェックインを行っていない可能性があり，その点は制限事項である．Foursquare などの位置情報ベース SNS でのチェックイン行動を調査した文献 [15], [16] によると，プライバシの懸念から自宅や学校など日常的に訪問する場所や病院ではチェックインしないユーザが多い [15]，自宅から近い場所ではチェックインしないユーザが多い [16] と報告されている．逆にあまり訪問したことがない場所，友人に格好いい（cool）と思われる場所ではチェックインしやすい傾向があるとも報告されている [15], [16]．これらのことから，日常行動の分析より観光やお出かけの行動分析に適していると考えられる．また，SNS のチェックインデータでは，ユーザに偏りがあると考えられるため，より偏りのない移動データでの検証は今後の課題である．4 章の評価において，神奈川県の入込観光客数の予測問題で評価を行うため，チェックインデータを神奈川県内のデータにフィルタリングした．ユーザ数は約 3 万 8 千で，チェックイン数は約 53 万件である．位置情報の粒度は，各チェックインが行われた市町村レベルに統一し，市町村間の移動を抽出した．政令指定都市では区レベルの情報まで取得可能であるが，3.4 節で評価に利用する入込観光客数のデータが市レベルであったため，粒度を合わせ市レベルに統一した．その後，各市町村をノードとし，移動があった市町村間にエッジを張ることで市町村のグラフを構築した．エッジには移動したユニークユーザ数による重みを設定する．

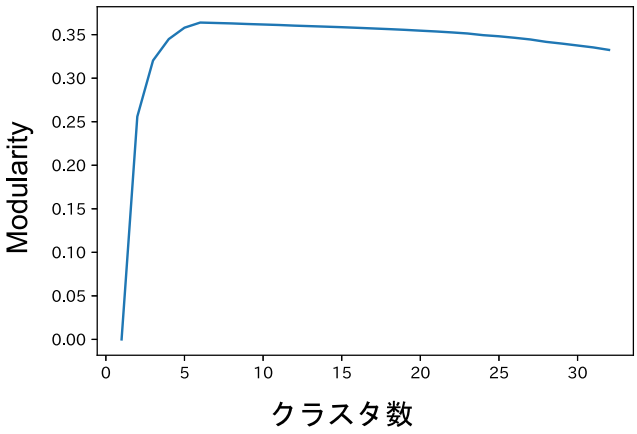

図 1 CNM Algorithm においてクラスタ数を変えたときの Modularity の推移

Fig. 1 Changes in modularity with different number of clusters using CNM Algorithm.

3.3 市町村のクラスタリング

グラフのクラスタリング手法は多数提案されている．本研究では，クラスタリング手法の違いによる結果への影響を考察するため Newman Algorithm [2] および Clauset-Newman-Moore（CNM）Algorithm [3] を利用してクラスタリングを行った．どちらのアルゴリズムもクラスタリングの質を Modularity [17] により評価する手法である．Modularity はクラスタ内とクラスタ間のエッジ数の割合に基づきクラスタリングの質を評価する指標である．Newman Algorithm では，まず各ノードをクラスタ（つまりノード数＝クラスタ数）と見なし Modularity を計算する．次に 2 つのクラスタを統合して Modularity が最も大きくなるクラスタを統合する．この操作を繰り返し全体が 1 つのクラスタになるまで Modularity を計算し，Modularity が最大となったクラスタを最終的なクラスタとする方法である．CNM Algorithm は Newman Algorithm をノード次数が小さい場合に計算量を削減できるようにしたアルゴリズムである．**図 1** に CNM Algorithm においてクラスタ数を変えたときの Modularity の推移のグラフを示す．この推移から Modularity が最大になるクラスタ数を 6 と決定した．

各アルゴリズムでの市町村のクラスタリング結果を**図 3**，**図 4** および**表 2**，**表 3** に示す．ノードの色がクラスタを示し，エッジの太さがエッジの重み（市町村間を移動したユニークユーザ数）を示す．**表 1** および**図 2** に神奈川県が定めている地域区分を示す[*1]．神奈川県が定めた地域区分と移動データに基づく地域区分を比較する．地域名称の決め方は，各アルゴリズムでの市町村のクラスタリング結果では，神奈川県が定めた地域区分を東西や南北に分けていることから，神奈川県が定めた地域区分の名称に東部，西部，南部，北部を追加し地域名とした．また，鎌倉市は知名

[*1] https://www.pref.kanagawa.jp/docs/ie2/cnt/f530001/p780102.html

© 2022 Information Processing Society of Japan

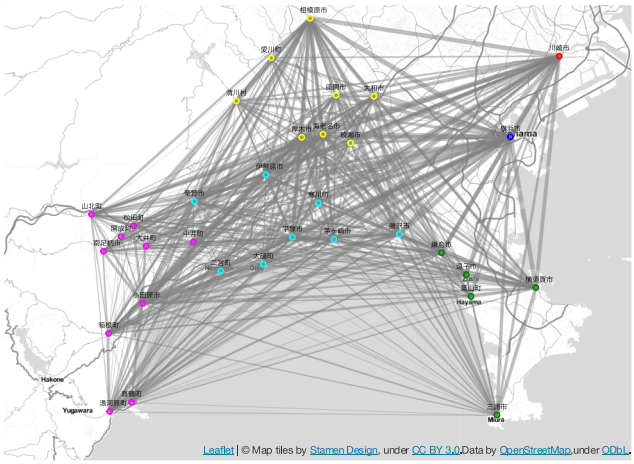

図 2 神奈川県が定めた地域区分
Fig. 2 Regions defined by Kanagawa Prefecture.

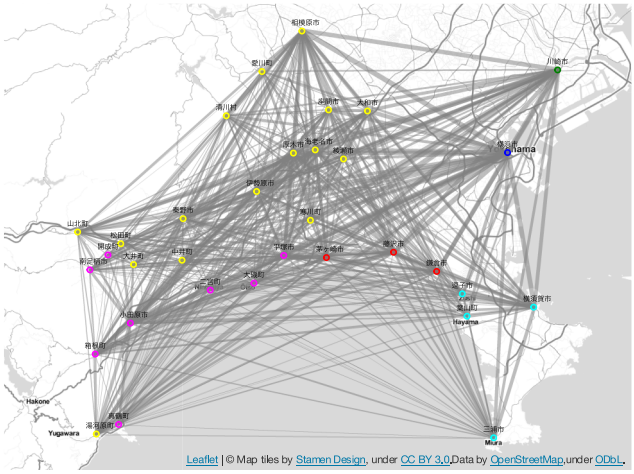

図 3 移動データに基づく市町村のクラスタ結果（Newman Algorithm）
Fig. 3 Regions determined by Newman Algorithm.

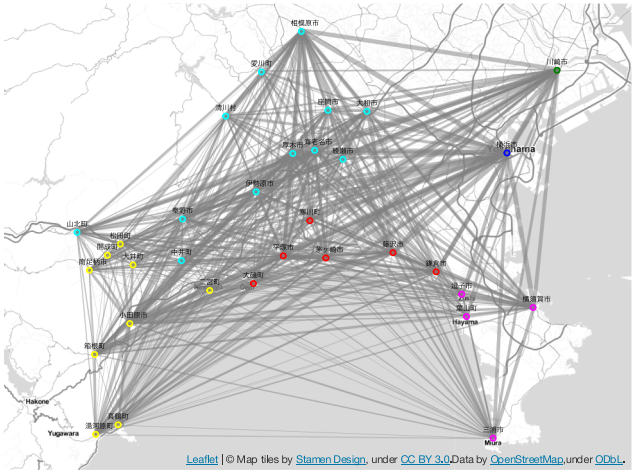

図 4 移動データに基づく市町村のクラスタ結果（CNM Algorithm）
Fig. 4 Regions determined by CNM Algorithm.

度が高いため単独で地域名に加えた．Newman Algorithm の場合は，平塚市，大磯町，二宮町が小田原市などと同一の地域と区分されている点が異なる．CNM Algorithm の場合は，山北町と二宮町が神奈川県が定めた地域区分と異なっている．また両方のアルゴリズムに共有して，鎌倉市が横須賀三浦地域ではなく，藤沢市や茅ヶ崎市と同様の湘南地域に分類されていることが特徴的である．

表 1 神奈川県が定めた地域区分
Table 1 Regions defined by Kanagawa Prefecture.

地区名	市町村名
横浜地域	横浜市
川崎地域	川崎市
横須賀三浦地域	横須賀市，鎌倉市，逗子市，三浦市，葉山町
県央地域	相模原市，厚木市，大和市，海老名市，座間市，綾瀬市，愛川町，清川村
湘南地域	平塚市，藤沢市，茅ヶ崎市，秦野市，伊勢原市，寒川町，大磯町，二宮町
県西地域	小田原市，南足柄市，中井町，大井町，松田町，山北町，開成町，箱根町，真鶴町，湯河原町

表 2 Newman Algorithm による地域区分
Table 2 Regions determined by Newman Algorithm.

地区名	市町村名
横浜地域	横浜市
川崎地域	川崎市
横須賀三浦地域	横須賀市，逗子市，三浦市，葉山町
湘南西部・県西南部地域	二宮町，南足柄市，大磯町，小田原市，平塚市，真鶴町，箱根町，開成町
湘南東部・鎌倉地域	茅ヶ崎市，藤沢市，鎌倉市
県央・県西北部地域	中井町，伊勢原市，厚木市，大井町，大和市，寒川町，山北町，座間市，愛川町，松田町，海老名市，清川村，湯河原町，相模原市，秦野市，綾瀬市

表 3 CNM Algorithm による地域区分
Table 3 Regions determined by CNM Algorithm.

地区名	市町村名
横浜地域	横浜市
川崎地域	川崎市
横須賀三浦地域	横須賀市，逗子市，三浦市，葉山町
湘南西部・県央地域	中井町，伊勢原市，厚木市，大和市，山北町，座間市，愛川町，海老名市，清川村，相模原市，秦野市，綾瀬市
湘南中東部・鎌倉地域	大磯町，寒川町，平塚市，茅ヶ崎市，藤沢市，鎌倉市
県西地域	二宮町，南足柄市，大井町，小田原市，松田町，湯河原町，真鶴町，箱根町，開成町

3.4 クラスタリングの評価

移動データに基づくクラスタリング結果をクラスタリングの活用先の指標で定量的に評価する．本研究では，移動データに基づく地域のクラスタリング結果を経済指標や公衆衛生のデータ分析などに活用することを想定している．そこで，一例として観光消費額の推移を利用して評価を行

© 2022 Information Processing Society of Japan

う．海や山などの観光資源が地域ごとに類似しているため，観光消費額は地域ごとに推移が似ていると想定できる．そのため，各地域区分に含まれる市町村の観光消費額の推移が類似しているほど，市町村のクラスタリングが適切に行えていると考えらえる．そこで，観光消費額の推移を行政で定めた地域区分と移動データに基づく地域区分で比較する．観光消費額は神奈川県の全市町村でデータが提供されているわけではないため，全市町村のデータが揃っている入込観光客数を利用する．ここで，入込観光客数と観光消費額は非常に強い相関があると考えられる．実際に神奈川県の 2018 年の入込観光客数データ[*2]で相関係数を計算したところ，相関係数 0.911，p 値 < 0.001 であった．そのため，入込観光客数の推移で評価した結果から観光消費額の推移についても同様の傾向があると考えられる．

具体的な計算方法を説明する．

(1) 市町村 i の入込観光客数の年間推移を pop_i とし，各月の入込観光客数を要素とする 12 次元のベクトルとする．各市町村の入込観光客数の最大値で割って正規化した正規化入込観光客数を $pop'_i = pop_i / \max(pop_i)$ とする．

(2) 2 つ以上の市町村を含むクラスタ $c \in C$ から任意の 2 市町村 $i, j \in c$ を抽出し pop'_i と pop'_j の距離 $D_{i,j}$ を計算する．距離 D の計算にはユークリッド距離を利用する．

(3) クラスタ c 内の全組合せで計算したユークリッド距離の平均値 D^c_{mean} を計算する．

$$D^c_{mean} = \frac{1}{|c|} \sum_{i,j \in c, i \neq j} D_{i,j}$$

(4) 全クラスタで D^c_{mean} の平均値を計算する．

$$D_{mean} = \frac{1}{|C|} \sum_{c \in C} D^c_{mean}$$

ここでのユークリッド距離は入込観光客数の年間推移の近さを表すため，観光客数の大小のパターンが類似しているほど小さい値となる．なお，代表的なクラスタリングアルゴリズムである k-means クラスタリング [18] で，クラスタの良し悪しを各データ点のクラスタ中心からの距離の平均で評価していることを参考に今回は平均値を利用した．平均値以外にも中央値や分散など他の統計量を利用することも考えられるため，その点は今後の課題とする．入込観光客数のデータは神奈川県が公表している 2018 年の入込観光客数データを利用した．以上の手順で距離を計算した結果を**表 4** に示す．なお，上記手順で距離を計算したため，クラスタに 1 つの市町村のみであった横浜市と川崎市は距離計算の対象から除外されている．人の移動に基づいて市町村をクラスタリングした方が分布間の距離が短いという結果であった．移動データでクラスタリングすることで入込観光客数の推移が似た市町村をまとめることができ，予測精度の向上が期待できる．

[*2] https://www.pref.kanagawa.jp/docs/ya3/cnt/f80022/p1202218.html（参照 2020-03-25）

表 4 入込観光客数推移の平均距離の比較
Table 4 Comparison of the average distance of the number of visitors.

クラスタリング方法	平均距離
県で定めた地方の分け方	1.323
Newman Algorithm	1.275
CNM Algorithm	1.238

4. 入込観光客数予測への応用

本章では，移動データに基づく地域のクラスタリング結果の応用事例として，入込観光客数予測問題で有効性の評価を行う．人の移動が予測対象の指標と因果関係がある場合に移動データに基づくクラスタリングを活用する効果が出ると考えられるため，人の移動が直接的に反映されると考えられる観光客数予測を応用として選定した．また，分析に使用したチェックインデータの性質上，お出かけや観光の分析に適していると考えられることも理由である．

4.1 入込観光客数予測問題

本節では，移動データに基づく地域のクラスタリング結果が予測精度の向上に寄与するかを検証する．予測モデルを構築する際，クラスタリングの結果を考慮する簡易な方法の 1 つとして，教師あり機械学習モデルに特徴量としてクラスタリングの結果を利用することが考えられる．そのため，ここでは教師あり機械学習モデルによる回帰モデルを検討する．一方，入込観光客数は時系列データのため，時系列モデルの自己回帰モデルをベースとした ARIMA，季節性を考慮した SARIMA などのモデルがよく用いられている [19]．そこで，これらの時系列モデルで考慮されている特徴を特徴量として予測モデルに利用する．時系列モデルの 1 つである季節調整モデルでは，観測値＝トレンド成分＋季節成分＋変動として要素を分解しモデル化を行っている [20]．本研究では，季節調整モデルを参考にそれらの要素の類似した特徴量を機械学習モデルの基本特徴量とし，移動データに基づく地域分けを追加の特徴量としてクラスタリングの結果を活用することの効果を検証する．

基本特徴量として以下の特徴量を利用する．

- 予測対象の市町村を示す one-hot ベクトル（$feat_{city}$）
- 予測対象の月を示す one-hot ベクトル（$feat_{month}$）
- 予測対象の市町村の過去直近 N カ月の入込観光客数（$feat_{trend}$）
- 予測対象の市町村の前年同月の入込観光客数（$feat_{seasonal}$）

上記の基本特徴量に，地域（クラスタ）を示す one-hot

© 2022 Information Processing Society of Japan

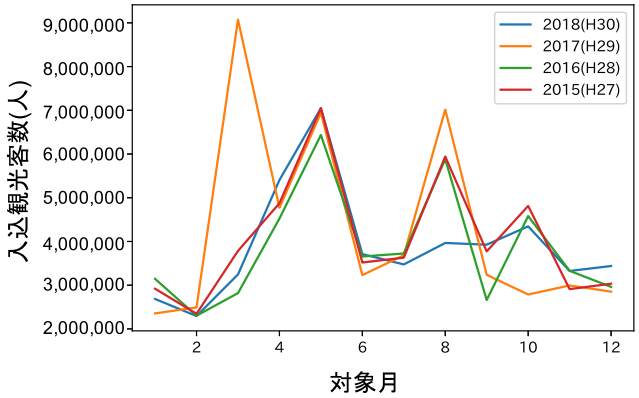

図 5 横浜市における 2015 年から 2018 年の入込観光客数の推移
Fig. 5 Transitions in the number of visitors to Yokohama from 2015 to 2018.

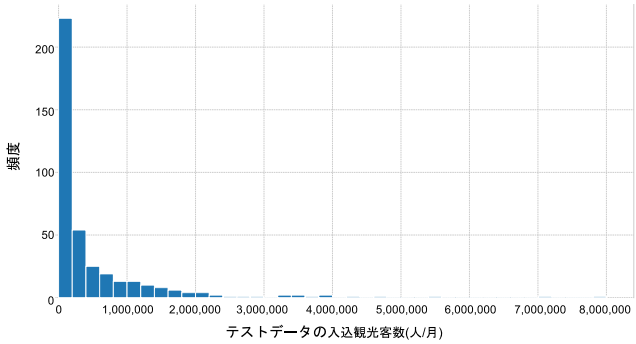

図 6 テストデータの入込観光客数の分布
Fig. 6 Distribution of the number of visitors for test data.

ベクトルを追加の特徴量（$feat_{cls}$）として加える．この追加特徴量には，行政で決めた地域区分または移動データに基づくクラスタを利用し，それぞれでの予測精度を比較することで移動に基づくクラスタの有用性を検証する．機械学習モデルには $feat_{city}$，$feat_{month}$，$feat_{trend}$，$feat_{seasonal}$，$feat_{cls}$ を連結したベクトルを特徴量として入力する．教師データとして各市町村の入込観光客数を利用する．機械学習モデルには様々なモデルが利用可能であるが，本研究では XGBoost 回帰モデル [21] を利用する．

4.2 評価条件

入込観光客数の予測には，前章で利用した移動データに基づく神奈川県の各市町村のクラスタリング結果と，2015 年から 2018 年の神奈川県入込観光客調査結果のデータを利用する[*3]．例として横浜市の入込観光客数の推移を**図 5** に示す．毎年大まかな傾向は同じであるが年によって大きく増減する月も存在する．機械学習モデルの学習データとして 2016 年と 2017 年の 2 年分を利用し，2018 年の 1 年分をテストデータとして利用する．

神奈川県は 33 市町村あるため，$33 \times 2 \times 12 = 792$ 件を学習データとして利用でき，$33 \times 12 = 396$ 件がテストデータとなる．テストデータにおける入込観光客数の分布を**図 6** に示す．図 6 より，実際の入込観光客数の分布には偏りがあることが分かる．テストデータの中央値を調べたところ，142,500 であった．そこで，全テストデータでの予測誤差と，正解データの値が 1 カ月あたり 15 万人未満のみ，および 1 カ月あたり 15 万人以上のデータのみでの予測誤差を評価指標とした．

追加特徴量には，神奈川県が定めた地域区分，Newman Algorithm によるクラスタリング，CNM Algorithm によるクラスタリングを利用し，それぞれの予測性能を比較することで，クラスタリング結果を活用することの有用性を評価する．予測性能の評価指標は，回帰問題の研究でよく用いられる二乗平均平方根誤差（RMSE：Root Mean Square Error）と平均絶対誤差（MAE：Mean Absolute Error）を用いる．XGBoost モデルのパラメータは，木の深さ（max_depth）および抽出される訓練データの割合（subsample）をグリッドサーチにより探索し，利用する各特徴量ごとに最適なパラメータを設定した．また，学習データが 792 件と比較的少ないため，評価の信頼性を向上させるため XGBoost モデルのシード値を 10 回変更して RMSE と MAE の平均値を計算した．

[*3] https://www.pref.kanagawa.jp/docs/ya3/cnt/f80022/p27746.html

表 5 入込観光客数予測の結果
Table 5 Forecast results of the number of visitors.

特徴量	RMSE	MAE
基本特徴量のみ	341,047	106,840
基本特徴量＋行政が決めた地域区分	339,622	106,766
基本特徴量＋Newman Algorithm	**320,655**	**103,072**
基本特徴量＋CNM Algorithm	332,633	104,441

4.3 評価結果と考察

全テストデータでの予測誤差の評価結果を**表 5** に示す．表中の太字は最も誤差が小さいことを示す．全テストデータでは，RMSE では Newman Algorithm を利用してクラスタリングした特徴量を利用した場合が最も誤差が小さく，行政が決めた地域区分と比較して誤差を 5.58%削減した．MAE でも Newman Algorithm を利用してクラスタリングした特徴量を利用した場合が最も誤差が小さく，行政が決めた地域区分と比較して誤差を 3.46%削減した．予測誤差に有意な差があるかを回帰分析に関する予測誤差に関する検定（Diebold-Mariano 検定）[22] により検証した．RMSE については，行政が決めた地域区分を利用した場合と Newman Algorithm を利用してクラスタリングした場合では，p 値 0.036 で有意水準 5%で有意な差があった．行政が決めた地域区分を利用した場合と CNM Algorithm を利用してクラスタリングした場合では，p 値 0.358 であり有意に差があるとはいえない結果であった．一方，MAE については，行政が決めた地域区分を利用した場合と Newman

© 2022 Information Processing Society of Japan

表 6 正解データが 15 万人/月未満での入込観光客数予測の結果
Table 6 Forecast results of the number of visitors with ground-truth data of less than 150,000 per month.

特徴量	RMSE	MAE
基本特徴量のみ	**24,629**	15,422
基本特徴量＋行政が決めた地域区分	25,686	15,393
基本特徴量＋ Newman Algorithm	24,780	**15,201**
基本特徴量＋ CNM Algorithm	24,922	15,286

Algorithm を利用してクラスタリングした場合では，p 値 0.031 で有意水準 5%で有意な差があった．行政が決めた地域区分を利用した場合と CNM Algorithm を利用してクラスタリングした場合では，p 値 0.127 であり有意に差があるとはいえない結果であった．

15 万人/月未満のデータのみの結果を**表 6** に示す．この条件では，RMSE では基本特徴量のみを用いた場合が最も誤差が小さく，MAE では Newman Algorithm を利用してクラスタリングした特徴量を利用した場合が最も誤差が小さかった．RMSE について，基本特徴量のみを用いた場合が最も誤差が小さかったため，基本特徴量のみを用いた場合と Newman Algorithm や CNM Algorithm を利用してクラスタリングした場合で予測誤差を Diebold-Mariano 検定したところ，p 値はそれぞれ 0.798，0.625 であり有意な差があるとはいえない結果であった．MAE については Newman Algorithm を利用してクラスタリングした特徴量を利用した場合が最も誤差が小さかったため，ベースラインとなる行政が決めた地域区分を用いた場合と提案手法（Newman Algorithm または CNM Algorithm を用いる）で Diebold-Mariano 検定したところ，p 値はそれぞれ 0.498，0.695 であり有意な差があるとはいえない結果であった．このような結果となった理由として考えられることは，入込観光客数が 15 万人/月未満の場合は各市町村の観光におけるオフシーズンであると考えられるが，クラスタリングに用いた移動データは 3.1 節で述べた半年間の市町村間の移動人数の総和を利用しており，人数が少ない時期の移動傾向をとらえられていない可能性がある．その対策としては，本研究で用いた移動データはユーザ数が限られたデータ（全ユーザに対する抽出率が低い）のため実施していないが，月別に移動データに基づく市町村グラフを構築することが考えられる．

次に，15 万人/月以上のデータのみの結果を**表 7** に示す．この条件でも，RMSE および MAE どちらの指標でも Newman Algorithm を使って移動データにより市町村をクラスタリングした結果を活用した場合が最も誤差が小さかった．RMSE について，行政が決めた地域区分を利用した場合と Newman Algorithm や CNM Algorithm を利用してクラスタリングした場合で予測誤差を Diebold-Mariano 検定したところ，p 値はそれぞれ 0.036，0.360 であり，Newman Algorithm を利用してクラスタリングした場合では有意水準 5%で有意な差があった．MAE については，行政が決めた地域区分を利用した場合と Newman Algorithm や CNM Algorithm を利用してクラスタリングした場合で Diebold-Mariano 検定したところ，p 値はそれぞれ 0.031，0.127 であり，こちらも Newman Algorithm を利用してクラスタリングした場合では有意水準 5%で有意な差があった．

表 7 正解データが 15 万人/月以上での入込観光客数予測の結果
Table 7 Forecast results of the number of visitors with ground-truth data of 150,000 or more per month.

特徴量	RMSE	MAE
基本特徴量のみ	486,611	202,028
基本特徴量＋行政が決めた地域区分	484,516	201,908
基本特徴量＋Newman Algorithm	**457,427**	**194,567**
基本特徴量＋CNM Algorithm	474,558	197,273

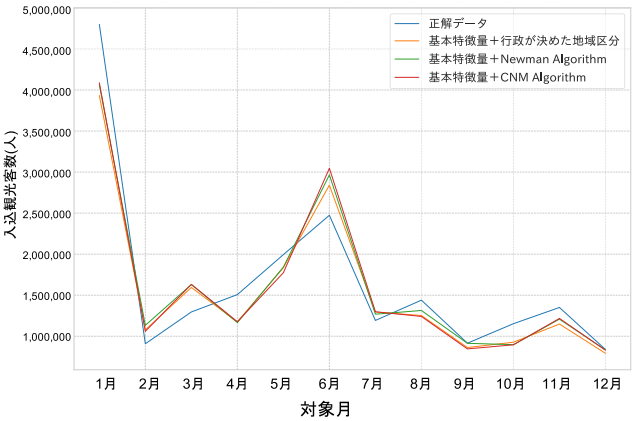

図 7 鎌倉市の入込観光客数の推移
Fig. 7 Transitions in the number of visitors to Kamakura City.

これらの結果から，移動データによる市町村のクラスタリングの結果は，入込観光客数予測問題において，特に観光客数が多い場合において有用であると考えられる．また，CNM Algorithm では有意な差を確認できなかったことから，複数のアルゴリズムでクラスタリングを行い，応用先によってクラスタリングのアルゴリズムを選定することが望ましいと考えられる．

入込観光客数が多い市町村の例として，鎌倉市の正解データと予測結果をプロットした図を示す（**図 7**）．鎌倉市では 1 月は初詣，6 月は紫陽花のスポットとして有名で観光客が多くなっている．6 月は行政が決めた地域区分を利用した場合が正解データに近いが，それ以外の月については移動データによるクラスタリングを利用した場合が正解データの推移に近いことが分かる．**表 8** に鎌倉市のみでの予測誤差の結果を示す．この結果では，RMSE，MAE のいずれにおいても Newman Algorithm の結果を特徴量として利用した方が誤差が小さい結果となった．

© 2022 Information Processing Society of Japan

表 8 鎌倉市の入込観光客数予測の結果
Table 8 Forecast results of the number of visitors regarding Kamakura City.

特徴量	RMSE	MAE
基本特徴量のみ	338,653	259,254
基本特徴量＋行政が決めた地域区分	326,575	250,791
基本特徴量＋Newman Algorithm	**315,282**	**240,671**
基本特徴量＋CNM Algorithm	325,031	257,990

5. おわりに

本研究では，人の移動データに基づいて市町村をクラスタリングしたデータを地域に関する統計指標の予測問題に活用することを提案した．Swarm の約 3 万 8 千ユーザ，約 53 万件のチェックインデータを利用し，市町村をノードとし，市町村間の移動があった箇所にエッジを設定することで市町村グラフを構築し，市町村をクラスタリングした．1 つめの評価として，各クラスタに含まれる市町村の観光消費額の推移が類似しているほど市町村のクラスタリングが適切に行えていると考え，観光消費額の推移の距離を行政が定めた地域区分と移動データに基づく地域区分で比較した．その結果，移動データに基づく地域区分の方が観光消費額が類似した市町村をまとめることができることを確認した．2 つめの評価として，移動に基づく地域のクラスタリングの結果を神奈川県の各市町村の入込観光客数の予測問題において有用性を確認した．行政が決めた地域区分を特徴量とした場合と比較して予測誤差を 5.58%削減できることを確認した．

今後の課題は，感染症など観光客数予測以外の地域に関する統計指標の予測での有用性を確認し，手法の一般性を評価することがあげられる．また，3.2 節や 4.3 節で述べたように，移動データに本研究では SNS のチェックインデータを利用したが，ユーザに偏りがあったり抽出率が低かったりするという制限があるため，より偏りがなく抽出率が高い位置情報データ（たとえば，携帯電話の基地局在圏情報に基づく人口データ）での評価を行いたい．図 3 および図 4 を比較すると，Newman Algorithm では同一クラスタの中の市町村が飛び地のように 1 つだけ離れた場所にあり，CNM Algorithm の方が市町村が塊として抽出できている．また，3.4 節の評価でも，平均距離が CNM Algorithm の方が短かったため，これらの結果からは CNM Algorithm の方が移動データのクラスタリングに適していると考えられる．しかしながら，4 章の予測問題の評価では，おおむね Newman Algorithm の方が誤差が小さかった．そのため，クラスタリング手法によって適した応用先が異なる可能性がある．今後，クラスタリングのアルゴリズムの特性と分析内容の関係についても明らかにしていきたい．

参考文献

[1] Foursquare: Swarm, Foursquare (online), available from ⟨https://www.swarmapp.com/⟩ (accessed 2020-04-28).
[2] Newman, M.E.: Finding community structure in networks using the eigenvectors of matrices, *Physical Review E*, Vol.74, No.3, p.036104 (2006).
[3] Clauset, A., Newman, M.E. and Moore, C.: Finding community structure in very large networks, *Physical Review E*, Vol.70, No.6, p.066111 (2004).
[4] Wang, Y., Yuan, N.J., Lian, D., Xu, L., Xie, X., Chen, E. and Rui, Y.: Regularity and conformity: Location prediction using heterogeneous mobility data, *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1275–1284 (2015).
[5] Zheng, Y., Liu, L., Wang, L. and Xie, X.: Learning transportation mode from raw gps data for geographic applications on the web, *Proc. 17th International Conference on World Wide Web*, pp.247–256 (2008).
[6] Nishida, K., Toda, H., Kurashima, T. and Suhara, Y.: Probabilistic identification of visited point-of-interest for personalized automatic check-in, *Proc. 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp.631–642 (2014).
[7] Ye, M., Yin, P., Lee, W.-C. and Lee, D.-L.: Exploiting Geographical Influence for Collaborative Point-of-interest Recommendation, *Proc. SIGIR '11*, pp.325–334 (2011).
[8] Yuan, Q., Cong, G., Ma, Z., Sun, A. and Thalmann, N.M.: Time-aware Point-of-interest Recommendation, *Proc. SIGIR '13*, pp.363–372 (2013).
[9] Kurashima, T., Iwata, T., Hoshide, T., Takaya, N. and Fujimura, K.: Geo Topic Model: Joint Modeling of User's Activity Area and Interests for Location Recommendation, *Proc. WSDM '13*, pp.375–384 (2013).
[10] Ochiai, K., Fukazawa, Y., Yamada, W., Manabe, H. and Matsuo, Y.: Gravity of Location-based Service: Analyzing the Effects for Mobility Pattern and Location Prediction, *Proc. International AAAI Conference on Web and Social Media* (*ICWSM*) (2020).
[11] Thiemann, C., Theis, F., Grady, D., Brune, R. and Brockmann, D.: The structure of borders in a small world, *PloS One*, Vol.5, No.11 (2010).
[12] Rinzivillo, S., Mainardi, S., Pezzoni, F., Coscia, M., Pedreschi, D. and Giannotti, F.: Discovering the geographical borders of human mobility, *KI-Künstliche Intelligenz*, Vol.26, No.3, pp.253–260 (2012).
[13] Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R. and Strogatz, S.H.: Redrawing the map of Great Britain from a network of human interactions, *PloS One*, Vol.5, No.12 (2010).
[14] Twitter：API リファレンス，Twitter（オンライン），入手先 ⟨https://developer.twitter.com/⟩（参照 2020-04-28）．
[15] Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J. and Zimmerman, J.: *I'm the Mayor of My House: Examining Why People Use Foursquare – A Social-Driven Location Sharing Application*, pp.2409–2418, Association for Computing Machinery (2011).
[16] Tasse, D., Liu, Z., Sciuto, A. and Hong, J.: State of the geotags: Motivations and recent changes, *Proc. International AAAI Conference on Web and Social Media*, Vol.11, No.1 (2017).
[17] Newman, M.E.: Modularity and community structure in networks, *Proc. National Academy of Sciences*, Vol.103, No.23, pp.8577–8582 (2006).
[18] MacQueen, J. et al.: Some methods for classification and

© 2022 Information Processing Society of Japan

analysis of multivariate observations, *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol.1, No.14, pp.281–297 (1967).
[19] 大井達雄：自己回帰実数和分移動平均モデルを使用した観光需要予測に関する考察，観光学，No.6, pp.1–7 (2012).
[20] 北川源四郎：時系列解析入門 (2005).
[21] Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system, *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785–794 (2016).
[22] Diebold, F.X. and Mariano, R.S.: Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, Vol.13, No.3, pp.253–263 (1995).

落合 桂一 （正会員）

2008 年千葉大学大学院博士前期課程修了．同年株式会社 NTT ドコモ入社．2017 年東京大学大学院工学系研究科博士後期課程修了．2020 年 8 月より東京大学特任助教．博士（工学）．SNS，位置情報，ヘルスケアデータやスマートフォンログ解析，FinTech 分野の研究開発に従事．ICWSM 2020 Best Paper Honorable Mentions 受賞．ACM，日本データベース学会各会員．

寺田 雅之 （正会員）

1995 年神戸大学大学院工学研究科修士課程修了．同年日本電信電話（株）入社．同社情報通信研究所，情報流通プラットフォーム研究所を経て，2003 年（株）NTT ドコモへ転籍．2008 年電気通信大学大学院電気通信研究科博士後期課程修了．博士（工学）．情報セキュリティ技術，プライバシ保護技術，大規模統計処理技術の研究開発に従事．2015 年度情報処理学会論文賞および山下記念研究賞受賞．2019 年度情報処理学会業績賞受賞．電子情報通信学会会員．

© 2022 Information Processing Society of Japan